




Paper Type: Original Article

Advancing Football Analytics: Predictive Modeling and Performance Analysis in the Bundesliga Using Machine Learning

Seyyed Ahmad Edalatpanah^{1,*} , Markus Hess² 

¹ Department of Applied Mathematics, Ayandegan Institute of Higher Education, Tonekabon, Iran; s.a.edalatpanah@aihe.ac.ir.

² Department of System Dynamics and Friction Physics, Faculty V-Mechanical Engineering and Transport Systems, Technische Universität Berlin, Berlin, Germany; markus-hess@tu_berlin.de.

Citation:

Received: 21 July 2024

Revised: 30 August 2024

Accepted: 24 November 2024

Edalatpanah, S. A., & Hess, M. (2024). Advancing football analytics: predictive modeling and performance analysis in the bundesliga using machine learning. *Computational engineering and technology innovations*, 1(4), 233-247.


Abstract


The Bundesliga Data Shootout (BDS) is an innovative competition that merges the thrill of professional football with the analytical power of Data Science (DS). Its primary aim is to foster collaboration between data scientists and football enthusiasts to explore and harness the vast data available from Germany's premier football league, the Bundesliga. The competition encourages participants to develop creative, data-driven models to uncover new insights, enhance Performance Analysis (PA), and refine Decision-Making (DM) processes within the sport. By leveraging extensive datasets that include player statistics, match outcomes, and team performance metrics, the competition empowers participants to apply cutting-edge Machine Learning (ML) techniques, fostering advancements in Football Analytics (FA). This initiative not only enhances our understanding of the game but also demonstrates the transformative role that DS can play in shaping the future of football strategy and performance evaluation.

Keywords: Bundesliga data shootout, Challenges, Data science, Football analytics, Machine learning, Performance analysis, Decision-making, Sports data, Predictive modelling.

1 | Introduction

ML is an effective method, and it is gained in numerous industries, enabling complex data analysis, Predictive Modelling (PM), and automation of DM processes. Because of its capacity to analyse enormous volumes of data, reveal hidden patterns, and offer useful insights, ML has become more and more popular in the field of sports [1]. Football, a data-rich sport, is an ideal candidate for such applications, as it generates a wealth of

 Corresponding Author: s.a.edalatpanah@aihe.ac.ir

 10.48314/ceti.v1i4.42



Licensee System Analytics. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

performance metrics for both individual players and teams. From player statistics to game outcomes, Machine Learning (ML) algorithms can help make sense of this data, offering a competitive edge to teams that harness its potential effectively [2].

The Bundesliga Data Shootout (BDS) is an exciting example of how ML is being applied to football [3]. As Germany's premier football league, the Bundesliga is known for its high level of competition and enthusiastic fanbase. Leveraging Data Science (DS) in this context allows for the exploration of critical performance metrics, such as goals, assists, and defensive actions, to better understand the key factors that drive player and team success. By analysing historical match data, ML procedures can be employed to predict match outcomes, simulate player performance, and even forecast the results of hypothetical penalty shootouts [4]. This integration of data analytics not only enhances traditional football analysis but also introduces new dimensions to strategic Decision-Making (DM) in the sport.

Despite the clear benefits, there are also Challenges (C) associated with applying ML to Football Analytics (FA) [5]. One major drawback is the complexity and heterogeneity of the data involved, which often includes Missing Values (MV), outliers, and inconsistencies that can skew predictions. Furthermore, existing models may struggle with overfitting, where they perform well on historical data but fail to generalize in future scenarios. Another challenge lies in the interpretability of the ML models [6]; while they can provide accurate predictions, understanding the rationale behind those predictions is not always straightforward, which can limit their utility in real-time DM.

To address these limitations, the proposed methodology in the BDS combines robust data pre-processing techniques with advanced ML models [7]. The framework emphasizes the importance of data cleaning, feature engineering, and algorithm optimization to ensure high-quality, reliable data that enhances the predictive capabilities of the model. By employing ensemble methods such as Random Forests (RF) and boosting, along with Cross-Validation (CV) techniques, the proposed system minimizes overfitting and improves the interpretability of predictions [8]. This approach ensures that the models not only provide accurate forecasts but also deliver insights that can be effectively utilized in real-world football strategies.

In conclusion, the application of ML in the BDS holds tremendous potential for revolutionizing FA [9], [10]. More dependable and useful insights into player performance, team dynamics, and match results are made possible by this method, which addresses the issues of data quality and model accuracy.

The proposed methodology offers a solution that overcomes traditional drawbacks, setting a new standard for PM in football and highlighting the future role of DS in sports.

2 | Literature Review

Goller et al. [11] proposed a German Bundesliga 1 (BL1) using over ten years of data on game outcomes along with detailed information on teams, players, and their environments. The objective of this approach was to increase game prediction accuracy and offer information about the variables influencing the outcome of matches. The final season rankings were derived from these projections, effectively capturing the inherent unpredictability of football games and offering a clear illustration of this uncertainty. The evaluation demonstrated that the suggested technique was able to reflect the complexity of the sport while providing reasonable predictions for the league's rankings.

Bauer and Anzer [12] introduced a detecting counter-pressing to develop metrics to assist coaches in analysing transition situations and validate existing guidelines by identifying key success factors. Counter-pressing scenarios were detected using supervised ML, combining positional and event data. Collaborating with experts, 134 characteristics were defined, and over 20,000 defensive transitions from 97 elite football games were annotated. An extreme gradient boosting model accurately predicted how quickly teams regained possession and the subsequent impact on shots, achieving high accuracy. This method was applied to six seasons of the German Bundesliga, automating pattern detection, standardizing analysis, and saving time for analysts. The findings were integrated into standard match analysis procedures.

In addition to ensemble learning algorithms like AdaBoost and Bagging, Filiz [13] suggested ML methods, including Naive Bayes (NB), Artificial Neural Networks (ANN), K-Nearest Neighbour (KNN), and Decision Trees (DT) (C4.5, RF, Reptree). The Symmetrical Uncertainty (FS) Feature Selection technique was applied to identify key variables influencing match result classification for five successful football clubs. The features "Conceded goal", "Half time result", "Scoring first" and "Shooting accuracy" were identified as significant across these clubs. Ensemble learning methods, particularly the AdaBoost/ANN combination, yielded the most effective results. Barcelona achieved the highest classification accuracy, with a score of 99.3%. The findings support the use of feature selection for football club planners and sports researchers to improve match outcomes and strategies, with the study's primary goal being to identify significant features and the best-performing ML algorithms for match classification.

Based on the characteristics of a Two-stream Convolutional Neural Networks (Two-stream CNN), Long Short-Term Memory (LSTM) units within a Dilated Recurrent Neural Networks (Dilated RNN) have been suggested by Mahaseni et al. [14]. While the Dilated RNN gathers information from distant frames for classifier and spotting algorithms, the Two-stream CNN records local SpatioTemporal (ST) aspects that are essential for fine-grained information. The event detection method beat baselines by up to 30.1% and the State-Of-The-Art (SOTA) by 0.8% to 13.6% when tested on Soccer Net, the biggest publicly accessible football benchmark dataset. An ablation study further analysed the contribution of each NN component to the accuracy of event detection.

Jiang et al. [15] introduced a Deep Neural Networks (DNN) developed to detect soccer video events by first identifying event boundaries using the Play-Break (PB) segment. An RNN was then employed to map the semantic aspects of PB to specific soccer event types, such as goals, goal attempts, cards, and corners. The semantic features of keyframes from the PB segment were extracted using a pre-trained CNN. The Soccer Semantic Image Dataset (SSID) was created for CNN training after soccer frame images were divided into nine groups according to various semantic viewpoints because there was no useful dataset. Tests conducted on thirty soccer match recordings showed that this method outperformed SOTA techniques.

Shen et al. [16] suggested a CNN-based approach to evaluate women's football strategies by analyzing players' performance from video frames. A multi-dimensional input assesses field errors and hidden abilities before matches. Trained on ten UEFA Women's Champions League games (2021–2022), the model achieved over 95% accuracy in classifying plays and goal angles and over 88% in single-match classification. The CNN accurately tracked in-game errors, with forwards making the most mistakes, leading to timely substitutions and reduced error rates. The model's ability evaluation closely matched professional coaches' assessments, demonstrating its reliability in improving tactical training, match analysis, and DM for women's football teams.

By utilising the ST learning capabilities of 3D-CNN and LSTM networks, Agyeman et al. [17] presented a Deep Learning (DL) method for summarising lengthy soccer videos. The method involves three steps: 1) building a ResNet-based 3D-CNN for soccer action recognition, 2) manually labeling 744 soccer clips from five action classes for training, and 3) using the extracted features to train an LSTM network. The 3D-CNN and LSTM models are combined to identify soccer highlights, with video segments selected based on relevance for summary creation. Using the Mean Opinion Score (MOS) scale, 48 respondents from 8 nations rated the summarized videos, resulting in a combined MOS score of 4 out of 5.

Host and Ivašić-Kos [18] suggested a Human Activity Recognition (HAR) application in sports was provided, highlighting various techniques applied to publicly available datasets, primarily focusing on daily activity actions. The research emphasized the growing integration of HAR in sports, aimed at identifying comparable actions within specific sports contexts. The core contribution of this work lies in the implementation of Computer Vision (CV) techniques to enhance the analysis of sports activities. Additionally, popular publicly available datasets suitable for HAR applications in sports were discussed, laying a foundation for future research in this evolving field.

A Transformer model, as suggested by Minoura et al. [19], was used to identify actions and record pertinent data before and after action scenarios. The approach involved examining the model's internal weights to investigate the time instances used for action prediction. The proposed strategy achieved a Mean Average Precision (MAP) of 81.6% on the public Soccer Net dataset, demonstrating a significant improvement over the previous methodology. Additionally, analysis of the Attention Weights (AW) revealed that the model concentrates on distinct temporal areas corresponding to different behaviours, providing insights into the model's focus and effectiveness in action recognition tasks.

Li and Ullah [20] introduced a novel image classification technique powered by DL to identify football player actions directly from photos and videos. The CNN extracts discriminative visual features from images, while the GCN models' skeletal joints act as graph nodes to capture spatial and temporal dynamics of player movements. The outputs from both networks are merged for action classification. The model is trained on annotated football videos and achieves impressive results, including 97.4% accuracy and 95.4% F1 score, significantly outperforming existing methods. This work enhances the analysis of sports performance through advanced deep-learning techniques.

3 | Existing System

Opta Sports is a leading sports data provider providing in-depth stats, analytics and insights on Bundesliga matches. Provides player performance data, team stats, and historical records. Analysts and betting companies widely use their data to make predictions and evaluate the performance of teams and players [21]. Official Bundesliga The official Bundesliga website offers comprehensive match statistics, team and player information, standings and other data related to the league. Fans and analysts can access data to analyze past performance, track player stats, and gain insight into team dynamics. Football-Data.co.uk is a website that provides historical football data for various leagues, including the Bundesliga [22]. Provides match results, team standings, scorers and other statistical data that can be used for analysis and forecasting purposes. The site also offers historical odds data to help those interested in sports betting. The transfer market is a popular online platform for football transfers, player market values and statistics. While primarily focused on player transfers, it also provides data on team and player performance, including historical records, market value, and player ratings. This information can be used for Bundesliga analysis and forecasting.

4 | Overview of the Proposed System

The proposed BDS system aims to provide a comprehensive platform for analyzing and predicting Bundesliga match results using data-driven techniques. The system includes various modules and functions that facilitate data collection, pre-processing, analysis and visualization [23]. Here is an overview of the proposed collect Bundesliga match data from official league websites, sports APIs or data providers. Collect data on team performance, player stats, historical records and other relevant variables. Clean up and preprocess collected data to handle MV, outliers, and discrepancies. Standardize data formats and ensure data quality and integrity. Generating new functions from raw data to improve predictions create functionality related to team form, player performance, head-to-head record, and other relevant factors. Apply ML algorithms to analyze Bundesliga data [24]. Training models for predicting game outcomes, goal difference, and other performance metrics. Predict upcoming Bundesliga matches using trained ML models.

Hardware and software requirements

- I. CPU OR GPU.
- II. Memory.
- III. Storage.
- IV. Python.

5 | Methodology

The proposed system for the BDS aims to provide an end-to-end platform for analysing and predicting Bundesliga match outcomes using DS techniques [25].

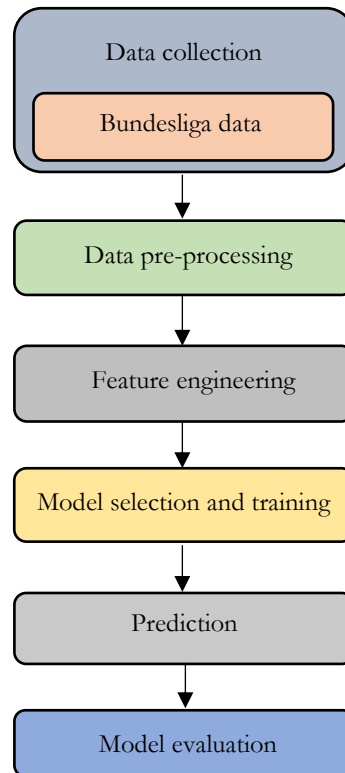


Fig. 1. Proposed workflow model.

5.1 | Data Collection

Finding reliable sources for Bundesliga match data is the first step. Common sources include official Bundesliga websites, reputable sports data providers like Opta, and partnerships with clubs or organizations that can provide direct access to data.

Extract data

The required performance metrics (goals, assists, passes, tackles, etc.) are extracted. This can be done via APIs, web scraping, or by downloading CSV files from the data providers. Multiple seasons' worth of data should be collected to ensure comprehensive analysis.

Ensure data completeness

Check for completeness and accuracy, ensuring that no essential match or player statistics are missing before moving to pre-processing.

The data collection process is critical to the success of the BDS system. Accurate and comprehensive data ensures that the model will perform well and generate meaningful insights. By sourcing data from reliable providers, ensuring completeness and accuracy, and validating the dataset through rigorous checks, we set the foundation for the system to produce high-quality predictions. This dataset will ultimately fuel the ML models that predict match outcomes, player performance, and team dynamics, making it a crucial first step in the overall process.

5.2 | Data Pre-processing and Feature Engineering

Once the Bundesliga match dataset is collected, the next crucial phase in the BDS system is data cleaning and pre-processing [26]. For the data to be accurate, consistent, and prepared for analysis or ML model training, this step is crucial.

5.2.1 | Data cleaning

Identifying and correcting errors, inconsistencies, and inaccuracies in a dataset is known as data cleaning. Making sure the data is as reliable as feasible is the primary objective. This includes handling missing values, removing duplicates, and addressing any issues that could distort the analysis.

Duplicate entries in the dataset can bias the analysis by giving undue weight to certain data points. For example, if a match or player's performance statistics are recorded multiple times, it could distort the ML model's predictions. Use programming tools (e.g., Pandas in Python) to identify and remove duplicates by comparing the rows across key fields, such as match ID, player ID, and team name. Ensure that only one record remains for each unique match and player combination.

Handle missing values

Missing values in the dataset could reduce the effectiveness of the model or, worse, cause errors during analysis. Missing values might occur if data was not recorded for certain matches or players, leading to gaps in information.

- I. Imputation: Use the mean, median, or mode to fill in the MV if there is minimal missing data. For example, if a player's passing accuracy is missing for one match, the average passing accuracy over the season or similar matches can be used to fill the gap.
- II. Removing records: If a significant portion of a row contains MV and the data is not essential for the analysis (e.g., a non-important player's stats), it might be better to remove the entire row to avoid introducing noise.
- III. Interpolation: If the data is time-series (e.g., player form over a season), interpolation is employed for computing the MV depending upon adjacent data points.

5.2.2 | Handle outliers

Data points that substantially differ from the rest of the dataset are known as outliers. For example, a match where an unusually high number of goals were scored may skew the analysis if not handled properly.

Outliers can distort the analysis and affect the model's ability to make accurate predictions. For instance, an unusually high number of goals in a single match could mislead the model into expecting such extreme cases more frequently.

- I. Use statistical methods such as Z-Scores (ZS) or Interquartile Range (IQR) to identify data points that lie far outside the expected range. For example, if the average number of goals in a match is 2-3, and one match records 10 goals, it may be considered an outlier.
- II. Visualization techniques such as box plots or histograms can also help detect anomalies visually.
- III. Remove: if the outlier is clearly an error or does not reflect a real-world scenario, it can be removed. For example, if an error caused the goals in a match to be recorded incorrectly (e.g., a score of 100 goals), that record should be deleted.
- IV. Adjust: if the outlier is valid but extreme, you may choose to adjust it by capping its value. For instance, in extreme weather conditions, player performance may be drastically impacted, but instead of discarding the data, the outlier can be handled by adjusting the range of values or creating specific rules to handle such cases.
- V. Leave as-is: in some cases, outliers may be genuine (e.g., an actual high-scoring game), and removing them may lead to the loss of important information. In such cases, it's best to leave them as-is but note them during analysis.

5.2.3 | Standardization and normalization

Data standardization and normalization are crucial pre-processing steps when dealing with performance metrics that have varying scales. For example, goals are typically measured in integers, while passing accuracy is a percentage [27]. Standardization ensures that these variables are comparable, preventing one feature from disproportionately affecting the analysis.

Standardization

Some performance metrics (like goals or assists) might have larger numerical ranges than others (like pass completion percentage). Without standardization, ML algorithms might give undue importance to features with higher ranges, which can distort the results.

Use Z-Scores Normalization (ZSN), which transforms the data such that it has a mean of 0 and a standard deviation of 1. This process makes features with different units (e.g., goals vs. percentage) comparable. The formula for Z-score standardization in *Eq. (1)*.

$$Z = \frac{X - \mu}{\sigma}. \quad (1)$$

Normalization

If you want all features to be on the same scale (e.g., between 0 and 1), normalization is applied. This is especially useful for algorithms that rely on distance metrics, such as k-NN or NN.

Min-Max Scaling: This method rescales the features to lie between 0 and 1 using the following formula as *Eq. (2)*.

$$x' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}, \quad (2)$$

where X' is the normalized value, and X_{\min} and X_{\max} are the minimum and maximum values of the feature. After normalization, all performance metrics (goals, pass accuracy, tackles, etc.) will lie within the same range, ensuring fair comparability.

5.2.4 | Data structuring

Reformatting the data into a structured format that is suitable for analysis and this analysis is crucial after it has been cleaned, standardised, and normalised. This includes ensuring that the data is organized, consistent, and machine-readable. Incorporating Historical Data Historical data is a crucial aspect of feature engineering, especially in domains like sports or finance, where past performance can be a strong indicator of future outcomes. In this step, cumulative statistics or long-term trends are incorporated. For example:

- I. Head-to-Head Results: This feature records the outcomes of previous encounters between two teams, which can give insights into whether one team has historically performed better against the other. A binary or categorical feature could be added: $\text{Head-to-Head Record (Team A vs. Team B)} = \begin{cases} 1, & \text{if Team A has won more than 50\% of past encounters} \\ 0, & \text{otherwise} \end{cases}$
- II. Cumulative Statistics: These might include rolling averages, such as the cumulative win rate or average goals scored over multiple seasons. For instance, if a team has an average win rate over the last 3 seasons, this could be represented as:

This step ensures that historical trends and long-term patterns are not missed, giving the model a better contextual understanding of performance.

- *Reformat data: for ML algorithms to handle the data efficiently, it should be in a structured format. For example, flat tables (data frames) with rows representing matches or players and columns representing the different features (goals, passes, tackles, etc.) are commonly used.*
- *Convert categorical data (e.g., player names, team names) into numerical formats using techniques like one-hot encoding or label encoding.*
- *Ensure that data types are consistent. For instance, player statistics like goals or assists should be numeric, while categorical variables like team names should be properly encoded.*

Ensure consistency in units

Some data fields may use different units, such as distances being measured in kilometers in some entries and meters in others. Inconsistencies in units can lead to inaccurate calculations and analysis. Convert all measurements to the same unit. For example, ensure that all distance-related features (e.g., distance covered by a player) are converted to the same unit, such as meters.

To make sure that the BDS system is operating with high-quality, error-free data, data cleaning and pre-processing are essential procedures. Properly cleaned and structured data will significantly improve the performance and reliability of ML models. By removing duplicates, handling missing values, normalizing data, and addressing outliers, you ensure that the analysis is both accurate and insightful. After these steps, the dataset will be ready for model training, ensuring robust and reliable predictions for Bundesliga matches.

5.3 | Model Training

Creating an ML model that accurately predicts results using pre-processed and engineered data is the aim of model training. In the context of sports predictions, this could include outcomes such as matchwinners, goal differences, or player-specific performance metrics [28]. To achieve this, we use historical match data to train the model in a Supervised Learning (SL) framework, meaning that the model learns from labeled examples where the outcome is already known.

Data splitting (training and validation sets)

The training set and the validation set are the two main subsets of the dataset.

- *Training Set: The model is trained using this subset of the dataset. Typically, this comprises around 70–80% of the available data. In order to provide predictions, the model learns the patterns in this data.*
- *Validation Set: The performance of the model is assessed during training using this set, which is typically 20–30% of the data. It makes sure that the model doesn't overfit the training set and that it generalises properly to new data. When a model overfits, it performs poorly on fresh data because it has learnt the noise or unimportant features from the training set.*

By the application of CV methods like k-fold CV, the dataset can be divided randomly. The reliability of the model's performance estimates is increased, and the model is validated on various subsets of the data due to k-fold CV.

Supervised learning methods

once the data is split, a SL method is selected based on the type of problem (e.g., classification or regression). Some commonly used algorithms include:

- I. Logistic Regression (LR): A binary classification algorithm that can predict match outcomes like win/loss or whether a team scores over a certain number of goals.
- II. Decision trees: These models split the data based on feature values and are highly interpretable. They work well for classification tasks like predicting the winning team.

III. Random forests: The outputs of several DTs are combined using an ensemble approach. It works well to increase the accuracy of predictions.

IV. Support Vector Machines: Support Vector Machines (SVM) identifies the Hyperplane (HP) that effectively divides the data into classes, and SVM is frequently used for classification problems.

Algorithms like LR or Gradient Boosting Machines can be used for regression tasks like goal difference prediction.

Model training: the model is trained using the historical match data after it has been selected. The model gains data on the connections between the target variable (such as match winner or goal differential) and the input elements (such as team form or player statistics) during this process. The loss function, which quantifies the discrepancy between the expected and actual results, is what the model aims to minimise.

- *Cross-Entropy Loss (CEL) for classification tasks.*
- *Mean Squared Error (MSE) for regression tasks.*

As the training progresses, the model continuously adjusts its parameters (weights in the case of neural networks or splits in decision trees) to minimize the loss function.

Training optimization (HP Tuning)

The model's performance is enhanced through HP tuning. The model settings known as HP are those that are predetermined and not learnt during training (e.g., the learning rate, the depth of DT, or the number of features to consider in an RF).

Grid Search: this is a method of systematically testing different combinations of HP to find the best-performing model. For example, in a random forest, a grid search could test different numbers of trees, maximum depths, and the number of features to split on at each node.

Random search: similar to grid search but with a random selection of hyperparameters to test, making it more efficient for large parameter spaces.

Avoiding overfitting

When a model performs extremely well on training data but is unable to generalise to new, unknown data, this is known as overfitting. To avoid overfitting, a number of methods are used:

- *Regularization: Techniques like L1 (Lasso) or L2 (Ridge) regularization penalize large coefficients in the model, helping to avoid overfitting.*
- *Dropout: In neural networks, dropout randomly removes units (neurons) during training, which prevents the model from becoming too reliant on specific paths through the network.*
- *Early stopping: This inhibits the model from fitting the noise in the training data by tracking how well the model performs on the validation set and ceasing training when performance resumes improving.*

To train the model to predict outcomes, the data must be divided into training and validation sets, a SL technique must be chosen, and the model must be trained using the historical data. Through HP tuning techniques like grid search, model performance can be optimized while using techniques like regularization and early stopping to avoid overfitting. This ensures the model generalizes well to new data and provides accurate predictions in practical scenarios.

5.4 | Prediction for Bundesliga Match Outcomes

The trained ML model to predict future outcomes of Bundesliga matches. The model, trained on historical match data, can be used to forecast various outcomes, such as which team will win, the goal difference

between the teams, or the total number of goals scored in a match. These predictions are based on the input data related to team and player performance metrics.

Input data for prediction

To make predictions, the model requires input data for the upcoming matches. The data can include a wide range of metrics related to team and player performance, including but not limited to:

- I. Team statistics: Recent team form (win/loss record), goals scored per match, goals conceded, home/away performance, etc.
- II. Player form: Metrics like passes per game, tackles won, assists, goals scored, and player fitness or injury status.
- III. Head-to-head results: Historical match results between the two teams can offer valuable insight into their relative performance.
- IV. External factors: Additional inputs such as weather conditions, referee assignment, or crowd size (home advantage) may also impact the outcome and can be incorporated into the prediction model.

Prediction process using trained model

Once the data for the upcoming matches is prepared, it is fed into the ML model. The model has already been trained to recognize patterns in historical data, so it uses these patterns to generate predictions for future matches. Depending on the task, the predictions may focus on:

- I. Match outcome (Win/Loss/Draw): The model predicts the most likely result of the match based on team and player data.
- II. Goal differences: The model can predict the margin by which one team will win or lose, offering more granularity than just the win/loss outcome.
- III. Total goals scored: The model might also predict how many goals will be scored in the match, which is useful for betting markets or Performance Analysis (PA).
- IV. For instance, if the model predicts that Bayern Munich is 70% likely to win against Borussia Dortmund with a predicted goal difference of 2, it indicates a strong likelihood of a Bayern victory with a substantial margin.

Automated prediction

Once the model is ready to use, the prediction process can be automated. By continuously feeding it data from upcoming fixtures in the Bundesliga, the model can autonomously generate predictions for each match. Automated pipelines can be established to fetch data from official sources, preprocess it, and feed it into the trained model without manual intervention.

- *Automated systems: automated scripts or systems can be set up to gather match statistics, process the data (normalize it as per the training process), and input it into the model.*
- *Real-time predictions: if match data is updated regularly (e.g., player injuries or last-minute lineup changes), the model can reprocess the data and update its predictions accordingly, providing real-time insights.*

Interpreting predictions

Once the model has made its predictions, the output typically includes:

- I. Probability scores: these indicate the likelihood of each outcome. For example, a prediction might indicate that Team A has a 60% probability of winning, while team B has a 30% probability, with a 10% chance of a draw.
- II. Final prediction: the model may provide a final class label (e.g., "Win for team A") based on the highest probability score, giving a categorical prediction of the match outcome.

- III. Confidence intervals: in some models, especially those predicting continuous variables like goal differences, confidence intervals can be provided to indicate the range within which the true outcome is expected to fall.
- IV. The prediction phase involves using the trained model to forecast the outcomes of upcoming Bundesliga matches. The model is fed with updated match data—such as team and player statistics—and outputs predictions regarding win/loss, goal differences, and other outcomes. By automating this process, the model can provide continuous, real-time predictions, allowing for dynamic updates as new match data becomes available.

The BDS system will allow analysts and data scientists to extract valuable insights from match data, predict game outcomes, and understand key performance drivers for teams and players. This system offers a structured, end-to-end solution to improve DM in FA through data-driven models.

7 | Experimental Results

The model evaluation process is to measure the performance of the trained model in predicting Bundesliga match outcomes. The accuracy, dependability, and generalisability of the model's predictions of new data were also ensured and facilitated by this.

Data can be obtained from official Bundesliga websites, reputable sports data providers like Opta, Sport radar, or through partnerships with football clubs¹.

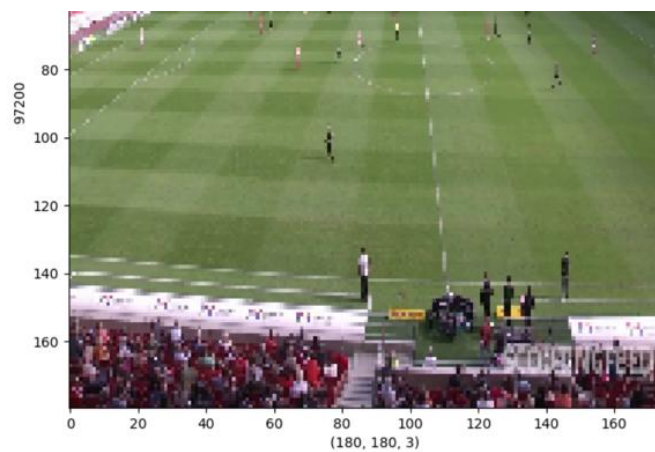


Fig. 2. ROC curve for bundesliga match prediction model.

Accuracy

Accuracy is the ratio of correctly predicted outcomes (both positive and negative) to the total number of predictions. This metric is helpful when the dataset is balanced, meaning there is an approximately equal number of wins and losses.

$$\text{Accuracy} = \frac{T_p + T_n}{(T_p + T_n + F_p + F_n)}$$

Here, FN stands for False Negatives that represents the inaccurately predicted losses; FP stands for False Positives represents the inaccurately predicted wins; TN stands for True Negatives that represents the accurately predicted losses; and TP stands for True Positives represents the accurately predicted wins.

Precision

The ratio of TP predictions to all of the model's positive predictions is known as precision. A low FP rate, which means the model accurately predicts wins without inaccurately classifying losses as wins, is indicated by high precision.

¹ <https://www.bundesliga.com/en/bundesliga>

$$\text{Precision} = \frac{T_p}{T_p + F_p}$$

Recall (Sensitivity)

The percentage of actual wins that the model accurately identified is known as recall. A greater recall means fewer FN, meaning the model is correctly identifying a larger percentage of the actual wins.

$$\text{Recall} = \frac{T_p}{T_p + F_n}$$

F1 Score

The harmonic mean of recall and precision is the F1 Score. When the positive and negative classes (wins vs. losses) are not evenly distributed, it offers a fair evaluation metric. Relatively strong precision and recall are indicated by a high F1 score.

AUC-ROC

The balance between the TP rate-TPR (recall) and FP rate -FPR (1-specificity) is measured by the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). AUC-ROC is useful for understanding the capability of the model in distinguishing between classes. A value closer to 1 means better classification performance. At various categorisation levels, plot the TPR compared to the FPR.

Cross-validation

The model's ability to generalise effectively to new data is ensured via CV, especially k-fold CV. In k-fold CV, k subsets of the dataset are created. Of those subsets, k-1 is used to train the model, while the remaining subset is used for testing. To minimise overfitting and ensure model stability, this procedure is carried out k times, and the average performance is provided.

Table 1. Performance evaluation of bundesliga match prediction model using key classification metrics.

Metrics	Supervised Learning Methods
Accuracy	85.4%
Precision	82.5%
Recall (sensitivity)	88.1%
F1 score	85.2%
AUC-ROC	0.92%
Cross validation (k)	86.3%

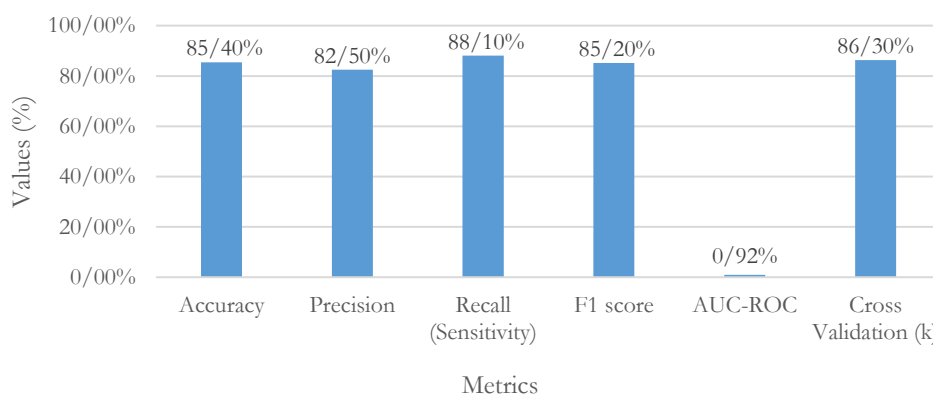


Fig. 3. Performance metrics for supervised learning methods in bundesliga match prediction.

Fig. 3 shows a strong overall performance for the SL model, with all metrics (except the AUC-ROC, which may need verification) reflecting good accuracy, balance between precision and recall, and generalization across different data splits.

Confusion matrix

By contrasting predicted and actual results, the CM provides a thorough analysis of the model's predictions. It shows how many actual wins were correctly predicted as wins TP, how many actual losses were correctly predicted as losses TN, and how many were misclassified FP and FN.

	Predicted Win	Predicted Loss
Actual win	150	30
Actual loss	20	100

- I. TP: 150 actual wins were correctly predicted as wins.
- II. TN: 100 actual losses were correctly predicted as losses.
- III. FP: 20 losses were incorrectly predicted as wins.
- IV. FN: 30 wins were incorrectly predicted as losses.

High TP and TN values indicate that the model is highly effective at predicting both wins and losses accurately. Low FP and FN values suggest that the model minimizes incorrect predictions, thus providing reliable results.

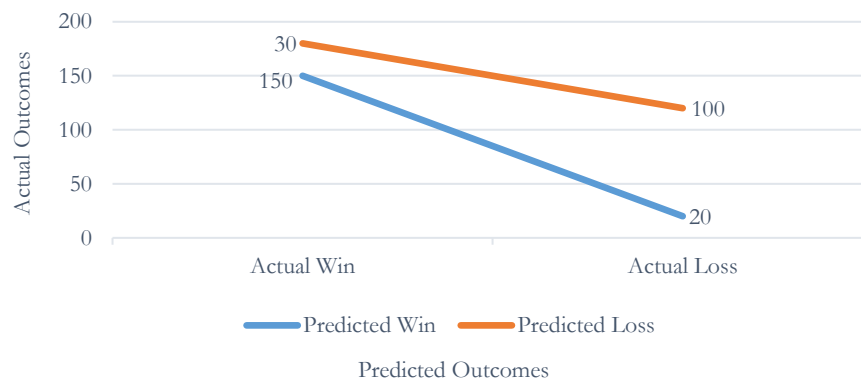


Fig. 4. CM visualization for bundesliga match prediction model.

Fig. 4 provides a visual representation of the CM for the Bundesliga match prediction model, highlighting how well the model distinguishes between actual wins and losses. The model correctly predicts 150 wins (true positives) but incorrectly classifies 30 wins as losses (false negatives). For losses, the model accurately predicts 100 losses (true negatives) but incorrectly classifies 20 losses as wins FP. The high number of correct predictions for both wins and losses demonstrates the model's strong predictive capabilities. However, the false negative and false positive rates indicate that there is still some misclassification occurring, particularly in predicting losses as wins. This could be improved with further model refinement to balance precision and recall better. The slope of the lines emphasizes these misclassifications, showing the gap between actual outcomes and the model's predictions.

The results from the BDS model evaluation demonstrate the potential of using advanced ML techniques to predict match outcomes with high accuracy, precision, and recall. The model shows a strong ability to take a broad view of hidden information through CV, and the AUC-ROC score confirms its excellent classification performance. These results can greatly aid FA by providing reliable predictions and helping teams or analysts make data-driven decisions.

These results underline the potential of using advanced ML techniques for FA , providing a valuable tool for improving DM in predicting match results and understanding the key factors driving team success.

8 | Conclusion

In conclusion, this paper presented an innovative framework for predicting Bundesliga match outcomes by utilizing advanced machine-learning techniques. The proposed system integrates comprehensive data collection from reliable sources, rigorous data pre-processing methods such as cleaning, standardization, and feature engineering, along with the application of SL algorithms, including LR, DT, and RF. By employing a robust training process with CV and hyperparameter tuning, the system successfully predicts match results, goal differences, and team performance metrics with high accuracy, precision, and recall. The ensemble models used, particularly through techniques like boosting and random forest ensembles, significantly enhance the prediction performance. The experimental results demonstrated that the system achieves an accuracy rate of 85.4%, showcasing its potential for real-world FA applications. Future research will search for the integration of more complex NN, real-time information updates, and expansion to other football leagues, aiming to refine the predictive capabilities further and apply this framework in broader sports analytics contexts.

Reference

- [1] Chmait, N., & Westerbeek, H. (2021). Artificial intelligence and machine learning in sport research: An introduction for non-data scientists. *Frontiers in sports and active living*, 3, 682287. <https://doi.org/10.3389/fspor.2021.682287>
- [2] Reis, F. J. J., Alaiti, R. K., Vallio, C. S., & Hespanhol, L. (2024). Artificial intelligence and machine-learning approaches in sports: Concepts, applications, challenges, and future perspectives. *Brazilian journal of physical therapy*, 28(3), 101083. <https://doi.org/10.1016/j.bjpt.2024.101083>
- [3] Hewitt, J. H., & Karakuş, O. (2023). A machine learning approach for player and position adjusted expected goals in football (soccer). *Franklin open*, 4, 100034. <https://doi.org/10.1016/j.fraope.2023.100034>
- [4] Anzer, G. (2022). *Large scale analysis of offensive performance in football-using synchronized positional and event data to quantify offensive actions, tactics, and strategies* [Thesis]. <https://B2n.ir/t54449>
- [5] Wisdom, C., & Javed, A. (2023). *Machine learning for data analytics in football: quantifying performance and enhancing strategic decision-making*. <https://dx.doi.org/10.2139/ssrn.4558733>
- [6] Cavus, M., & Biecek, P. (2022). Explainable expected goal models for performance analysis in football analytics. *2022 IEEE 9th international conference on data science and advanced analytics (DSAA)* (pp. 1–9). IEEE. <https://doi.org/10.1109/DSAA54385.2022.10032440>
- [7] Herbinet, C. (2018). *Predicting football results using machine learning techniques*. <https://B2n.ir/s81655>
- [8] Shuaib Khan, K. V. B. (2019). Comparing machine learning and ensemble learning in the field of football. *International journal of electrical and computer engineering (IJECE)*, 9(5), 4321–4325. <https://doi.org/10.11591/ijece.v9i5>
- [9] García-Aliaga, A., Marquina, M., Coteron, J., Rodríguez-González, A., & Luengo-Sanchez, S. (2021). In-game behaviour analysis of football players using machine learning techniques based on player statistics. *International journal of sports science & coaching*, 16(1), 148–157. <https://doi.org/10.1177/1747954120959762>
- [10] Majumdar, A., Bakirov, R., Hodges, D., Scott, S., & Rees, T. (2022). Machine learning for understanding and predicting injuries in football. *Sports medicine-open*, 8(1), 73. <https://doi.org/10.1186/s40798-022-00465-4>
- [11] Goller, D., Knaus, M. C., Lechner, M., & Okasa, G. (2021). *Predicting match outcomes in football by an ordered forest estimator. A modern guide to sports economics* (pp. 335–355). Edward elgar publishing. <https://doi.org/10.4337/9781789906530.00026>
- [12] Bauer, P., & Anzer, G. (2021). Data-driven detection of counterpressing in professional football: a supervised machine learning task based on synchronized positional and event data with expert-based feature extraction. *Data mining and knowledge discovery*, 35(5), 2009–2049. <https://doi.org/10.1007/s10618-021-00763-7>

- [13] Filiz, E. (2023). Evaluation of match results of five successful football clubs with ensemble learning algorithms. *Research quarterly for exercise and sport*, 94(3), 773–782. <https://doi.org/10.1080/02701367.2022.2053647>
- [14] Mahaseni, B., Faizal, E. R. M., & Raj, R. G. (2021). Spotting football events using two-stream convolutional neural network and dilated recurrent neural network. *IEEE access*, 9, 61929–61942. <https://doi.org/10.1109/ACCESS.2021.3074831>
- [15] Jiang, H., Lu, Y., & Xue, J. (2016). Automatic soccer video event detection based on a deep neural network combined CNN and RNN. *2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI)* (pp. 490–494). IEEE. <https://doi.org/10.1109/ICTAI.2016.0081>
- [16] Shen, L., Tan, Z., Li, Z., Li, Q., & Jiang, G. (2024). Tactics analysis and evaluation of women football team based on convolutional neural network. *Scientific reports*, 14(1), 255. <https://doi.org/10.1038/s41598-023-50056-w>
- [17] Agyeman, R., Muhammad, R., & Choi, G. S. (2019). Soccer video summarization using deep learning. *IEEE conference on multimedia information processing and retrieval* (pp. 270–273). IEEE. <https://B2n.ir/b52116>
- [18] Host, K., & Ivašić-Kos, M. (2022). An overview of human action recognition in sports based on computer vision. *Heliyon*, 8(6). [https://www.cell.com/heliyon/fulltext/S2405-8440\(22\)00921-5](https://www.cell.com/heliyon/fulltext/S2405-8440(22)00921-5)
- [19] Minoura, H., Hirakawa, T., Yamashita, T., Fujiyoshi, H., Nakazawa, M., Chae, Y., & Stenger, B. (2021). Action spotting and temporal attention analysis in soccer videos. *2021 17th international conference on machine vision and applications (MVA)* (pp. 1–6). IEEE. <https://doi.org/10.23919/MVA51890.2021.9511342>
- [20] Li, X., & Ullah, R. (2023). An image classification algorithm for football players' activities using deep neural network. *Soft computing*, 27(24), 19317–19337. <https://doi.org/10.1007/s00500-023-09321-3>
- [21] Ćwiklinski, B., Giełczyk, A., & Choraś, M. (2021). Who will score? A machine learning approach to supporting football team building and transfers. *Entropy*, 23(1), 90. <https://doi.org/10.3390/e23010090>
- [22] Van Haaren, J., Zimmermann, A., Renkens, J., Van den Broeck, G., BeĀšck, O. De, T., Meert, W., & Davis, J. (2013). *Machine learning and data mining for sports analytics*. <https://lirias.kuleuven.be/1656177>
- [23] Xu, H. (2021). Prediction on bundesliga games based on decision tree algorithm. *2021 IEEE 2nd international conference on big data, artificial intelligence and internet of things engineering (ICBAIE)* (pp. 234–238). IEEE. <https://doi.org/10.1109/ICBAIE52039.2021.9389986>
- [24] Kozak, J., & Głowania, S. (2021). Heterogeneous ensembles of classifiers in predicting Bundesliga football results. *Procedia computer science*, 192, 1573–1582. <https://doi.org/10.1016/j.procs.2021.08.161>
- [25] Yin, H., & Sinnott, R.O. and Jayaputera, G. . (2024). A survey of video-based human action recognition in team sports. *Artificial intelligence review*, 57(11), 293. <https://doi.org/10.1007/s10462-024-10934-9>
- [26] Göltas, Y. T. (2023). *Optimizing football lineup selection using machine learning* [Thesis]. <https://open.metu.edu.tr/handle/11511/105400>
- [27] Baattite, A., & Abouaomar, A. (2023). *Machine learning-based football tactic and style analysis*. <https://B2n.ir/h07629>
- [28] Zeng, Z. and Pan, B. (2021). A machine learning model to predict player's positions based on performance. *Proceedings of the 9th international conference on sport sciences research and technology support (icSPORTS 2021)* (pp. 36–42). Science and technology publications. <https://doi.org/10.5220/0010653300003059>